CNR-ITSC / IFOMIS
Workshop on Ontologies in Medicine
Rome, October 8-9, 2003

From biomedical language to biomedical knowledge

Uncovering relations expressed through reification and other linguistic phenomena

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Acknowledgments

- Songmao Zhang
- Anita Burgun
- Tom Rindflesch

Anatomy examples

- Foundational Model of Anatomy (C. Rosse & al.)
- GALEN (A. Rector & al.)

Outline

- Concepts and semantic relations
- Lexical phenomena representing semantic relations
- Applications
  - Building ontologies
  - Aligning ontologies
  - Validating ontologies

Introduction

Concepts and semantic relations

Concepts and semantic relations

- Knowledge representation paradigm
  - Concepts represent categories
    - Bacteria
    - Addison's disease
  - Semantic relations represent assertions
    - Propanolol *treats* Arrythmia
    - Adrenal gland *produces* Cortisol

Concepts vs. semantic relations

- Concept names may also embed
  - Assertions
    - Adrenal gland ↔ Adrenal gland *is a kind of* Gland
  - Predicates
    - Anti-arrythmia agent ↔ *treats* Arrythmia
    - Subdivision of heart ↔ *part of* Heart

From concept names to relations

- Semantic relations can be extracted from a combination of
  - Predicates embedded in concept names
  - Existing relations
- Examples of semantic relations extracted
  - Propanolol *treats* Arrythmia
    - Propanolol *isa* Anti-arrythmia agent
    - Anti-arrythmia agent ↔ *treats* Arrythmia
  - Cardiac chamber *part of* Heart
    - Cardiac chamber *isa* Subdivision of heart
    - Subdivision of heart ↔ *part of* Heart

Issues with explicit vs. implicit relations

- Redundancy
  - Some semantic relations may be represented both explicitly and implicitly
  - Maintenance issues
- Consistency
  - Implicit semantic relations may not always be represented explicitly
  - Issues:
    - In one ontology: Inconsistent representation
    - Across ontologies: Alignment issues

Linguistic phenomena
representing semantic relations

General framework

- Lexical semantics [Cruse, 1986]
  - Linguistic structures ↔ Semantic relations
- Knowledge acquisition from textual sources
  - General relations from general corpora
    - Hyponymy (*isa*), meronymy (*part of*)
      (e.g., from machine-readable dictionaries) [Dolan & al, 1993]
  - Specialized relations from the biomedical literature
    - e.g., *binds* from MEDLINE [Rindflesch & al, 2000]
  - Various relations from terminologies [Aussenac-Gilles & al., 1995]

Nominal modification

- [mod(adj|noun)+, head(noun)]$_{NP}$
  → Concept$_{NP}$ *isa* Concept$_{Head}$
- Adjective-Noun
  - Acute meningitis
    → Acute meningitis *isa* Meningitis
- Noun-Noun
  - Lung cancer
    → Lung cancer *isa* Cancer
- Domain independent

Nominal modification

- [mod(adj|noun)+, head(noun)]$_{NP}$
  → Concept$_{Mod}$ *rel* Concept$_{Head}$
- Mod-Head
  - Lung cancer
    → Lung *location of* Cancer
  - Viral infection
    → Virus *causes* Infection
- Domain dependent

Reification

- *Part of*
  - Component of X: Finger *isa* Component of hand ↔ Finger *part of* Hand
  - Subdivision of X: Cardiac chamber *isa* Subdivision of heart ↔ Cardiac chamber *part of* Heart
  - Organ component of X: Cardiac sphincter *isa* Organ component of stomach ↔ Cardiac sphincter *part of* Stomach
- *Branch of*
  - Branch of X: Sural nerve *isa* Branch of tibial nerve ↔ Sural nerve *branch of* Tibial nerve
- Other reified relations (function)
  - Iron transporter ↔ *carries* Iron (e.g., Ferritin) [Burgun & al, 2002]
  - Angiotensin-Converting Enzyme (ACE) inhibitor ↔ *inhibits* ACE (e.g., Captopril)

Prepositional attachment

- [head(noun), [prep(of), head(noun)]$_{PP}$]$_{NP}$
  → Concept$_{NP}$ *part of* Concept$_{PP}$
  - Muscle of pelvis → Muscle of pelvis *part of* pelvis
  - Nail of third toe → Nail of third toe *part of* third toe
  - Base of$_1$ phalanx of$_2$ middle finger
    → B of P of MF *part of* P of MF (of$_1$)
    → B of P of MF *part of* MF (of$_2$)
- Other prepositions
  - Urine test for glucose ↔ Urine test *analyzes* Glucose
- Domain dependent

Other phenomena

- Lexico-syntactic patterns for hyponymic relations
  - Appositives
    - Captopril, an ACE inhibitor, is used for …
  - Other patterns [Hearst & al, 1992]
    - **Such** ACE inhibitors **as** captopril are used for …
    - ACE inhibitors **including** captopril and enalapril are …
    - ACE inhibitors, **especially** captopril, are …
    - […]
- Relatively rare in concept names

Limitations (Anatomy)

[Zhang & al., 2003]

- Mostly unambiguous within a given subdomain
- Exceptions
  - Carotid body $\not\Rightarrow$ Carotid body *isa* Body
  - Groove for arch of aorta
    $\not\Rightarrow$ Groove for arch of aorta *part of* aorta

Applications

Applications

- Building ontologies
  - Acquire relations
  - Extend existing ontologies
- Aligning ontologies
  - Make knowledge explicit in both ontologies
- Validating ontologies
  - Compare existing relations to acquired relations
  - Compare existing concept names to potential concept names

Application 1 Building ontologies

[Bodenreider & al, 2001]

- Acquire terms from a corpus
  - MEDLINE
- Relate these terms to existing terms in UMLS
  - Adjectival modification (T_n = adj + T_o)
- Semantic relation: Hyponymy

Principles

A phrase from MEDLINE becomes a candidate term in the Metathesaurus if:

- *Condition 1: A demodified term created from this phrase is found in the Metathesaurus*

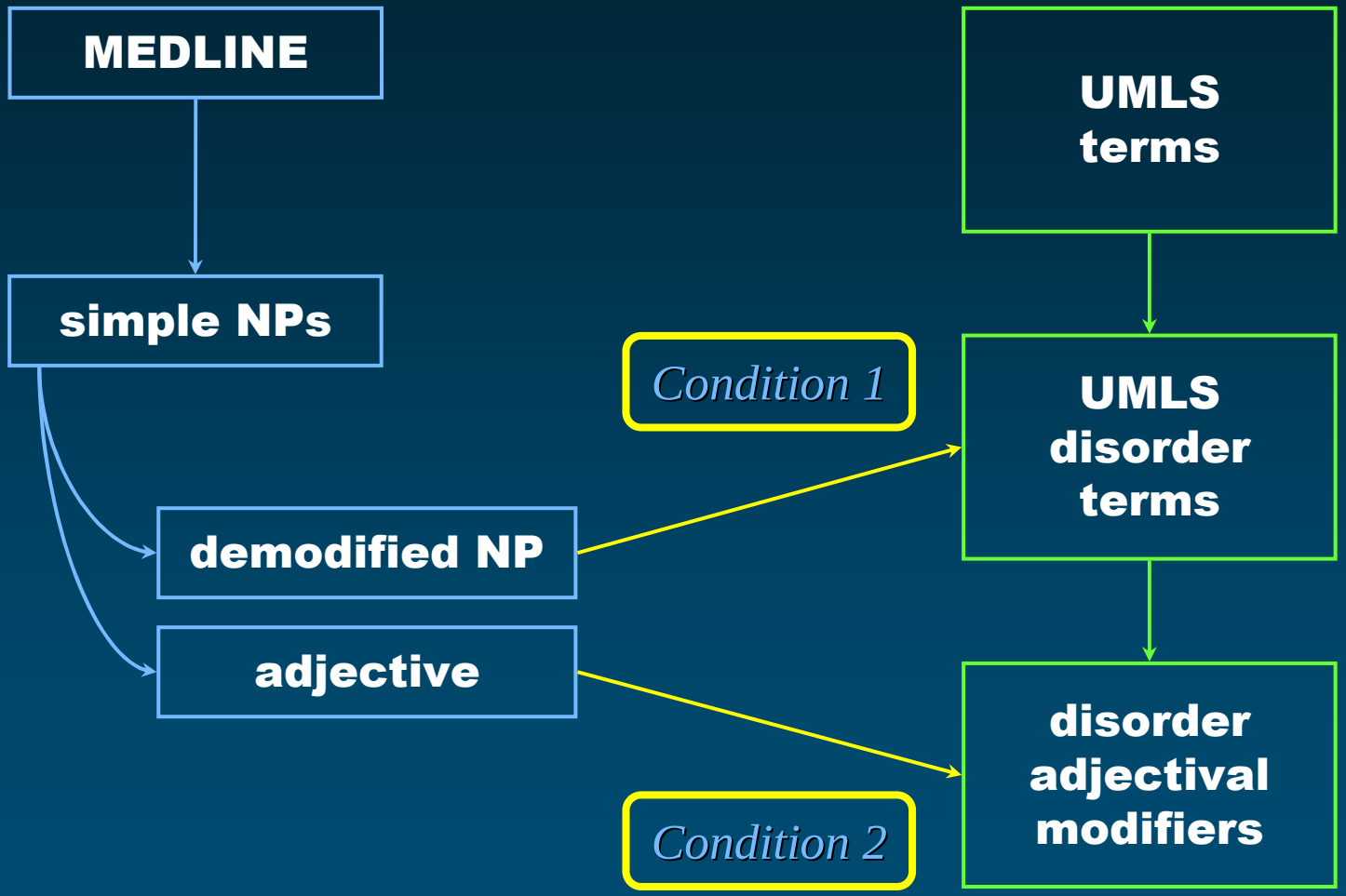
and

- *Condition 2: The modifiers removed from the MEDLINE phrase also modify existing terms from the Metathesaurus, for a given semantic category*

Example

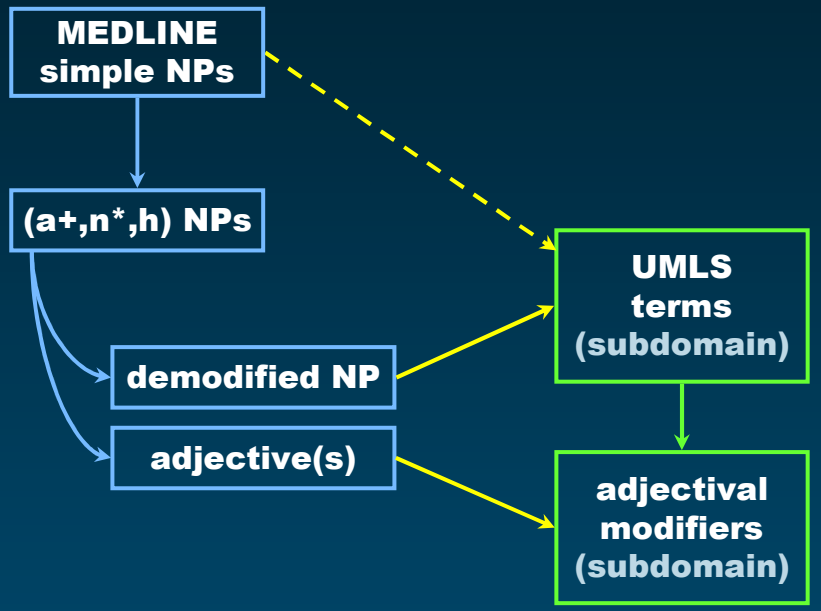
pancreatic bronchogenic cyst

= *pancreatic* + bronchogenic cyst



Results Quantitative

- 3 M “simple” MEDLINE NPs
- 21,000 already in the Metathesaurus (eliminated)
- 1.3 M (adj+, noun*, head) NPs
- 1.6 M demodified terms
- 125,464 candidate terms



Results Evaluation

- Limited evaluation
- 1000 candidate terms, randomly selected
- 1000 pairs (candidate term, Metathesaurus concept)
- Manual review
- Relevance of the association
  - 83% relevant
  - 3% more or less relevant

severe ocular inflammatory disease → disease

  - 14% not relevant

appropriate aid → AID - Artificial insemination by donor

Limitations

- Limited review
- Causes for non-relevant associations
  - Acronyms

    appropriate aid → AID - Artificial insemination by donor
  - Inaccurate POS tagging

    controlling stress → stress
  - Inaccurate mapping

    urinary protein → protein [measurement]

Application 2 Aligning ontologies

- 2 ontologies of anatomy [Zhang & al, 2003]
  - Foundational Model of Anatomy
  - GALEN
- Identify equivalent concepts
  - Lexical similarity
  - Structural similarity (hierarchical relations)
- Impaired by equivalent relations represented differently in the 2 ontologies (implicit/explicit)
- Knowledge augmentation

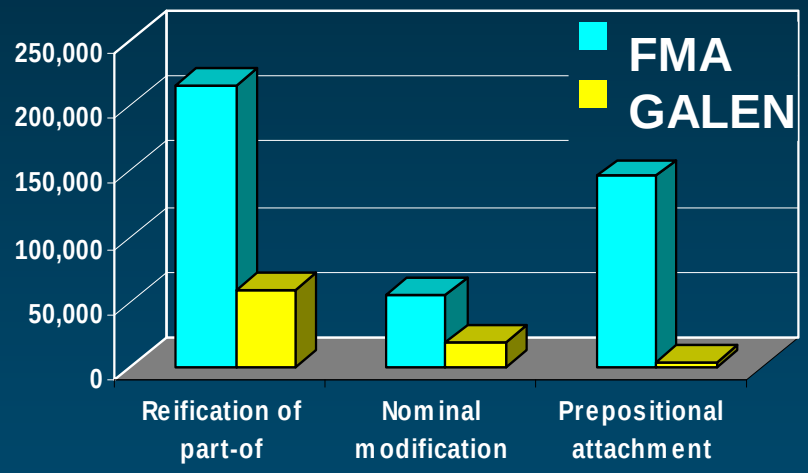
NLM

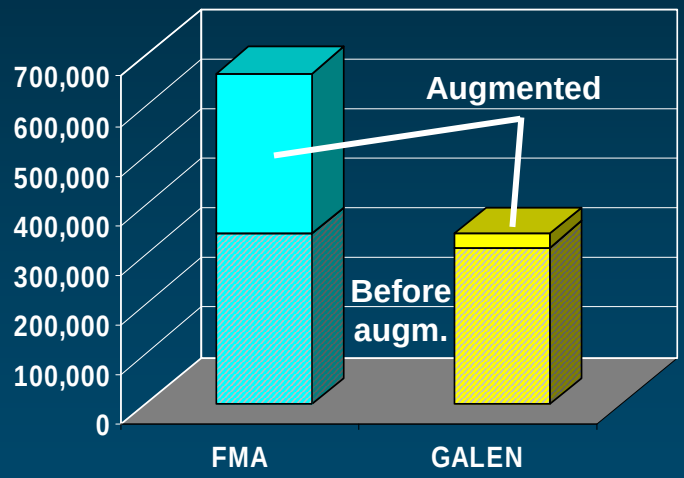
Augmentation methods

- *isa*
  - Nominal modification
- *part of*
  - Prepositional attachment (of)
  - Reification (subdivision of, component of, …)
- *branch of*
  - Reification (branch of)

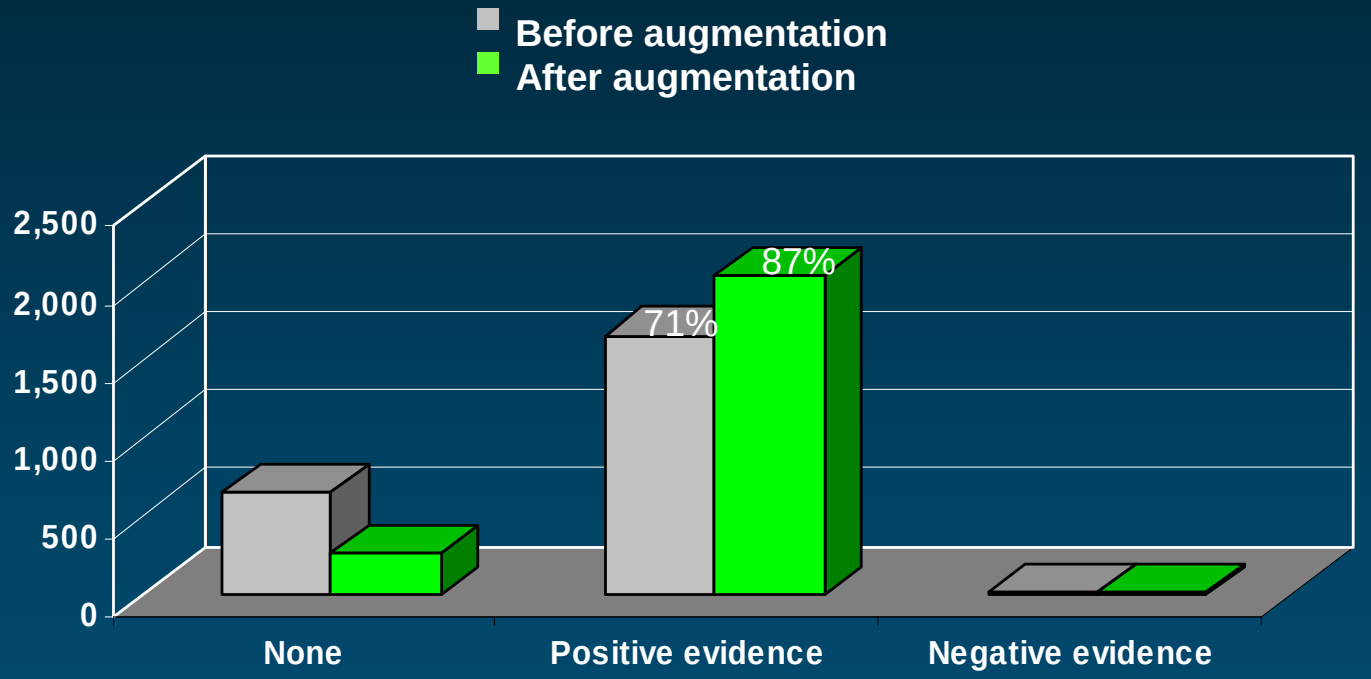
[Zhang & al., 2003]

Results


250,000
200,000
150,000
100,000
50,000
0
FMA
GALEN
Reification of part-of
Nominal modification
Prepositional attachment


700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
Augmented
Before augm.
FMA
GALEN

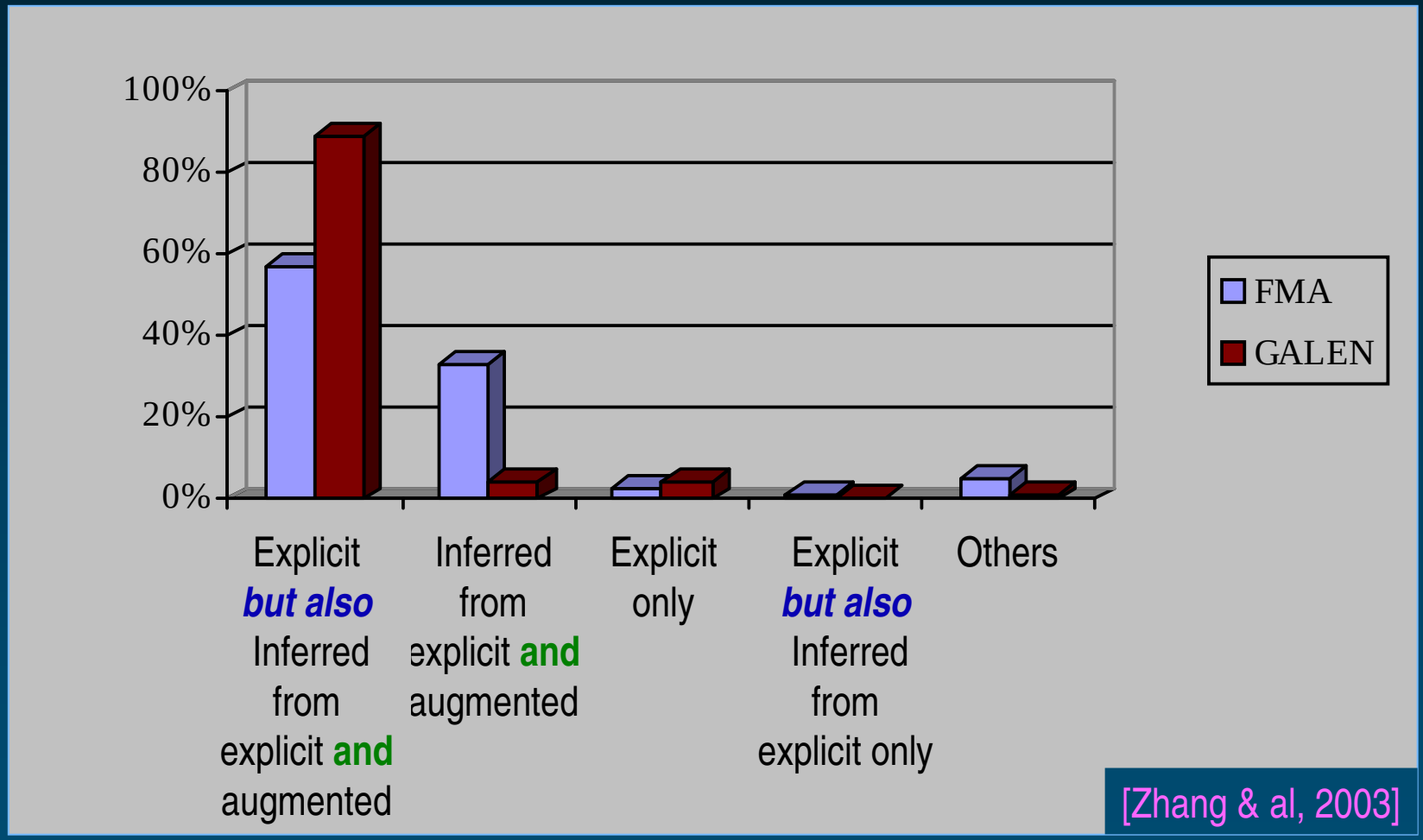
Evaluation


Before augmentation
After augmentation
2,500
2,000
1,500
1,000
500
0
71%
87%
None
Positive evidence
Negative evidence

Application 3 Validating ontologies

- Relations embedded in concept names are expected to be represented explicitly
- Relations generated by knowledge augmentation are expected to be consistent with explicit relations (within/across ontologies)

Explicit vs. implicit relations



[Zhang & al, 2003]

Inconsistencies revealed by augmentation

- Within ontologies [Zhang & al, 2003]
  - Internal spermatic fascia *isa* Organ component of internal spermatic fascia
  - Conflict
    - Explicit: Apex of urinary bladder *has part* Urinary bladder
    - Augmented: Apex of urinary bladder *part of* Urinary bladder (from Apex of urinary bladder *isa* Subdivision of UB)
- Across ontologies
  - FMA: Shoulder *part of* Pectoral girdle
  - GALEN: Shoulder *has part* Pectoral girdle

Other issues revealed by augmentation

[Bodenreider & al, 2002]

- Consistency of children in medical terminologies
- Based on adjectival modification
- Compare
  - Existing relations to expected relations
  - Existing concept names to potential concept names

Methods Co-occurrence of modifiers

primary lacrymal atrophy
secondary lacrymal atrophy
primary amyloidosis
secondary amyloidosis

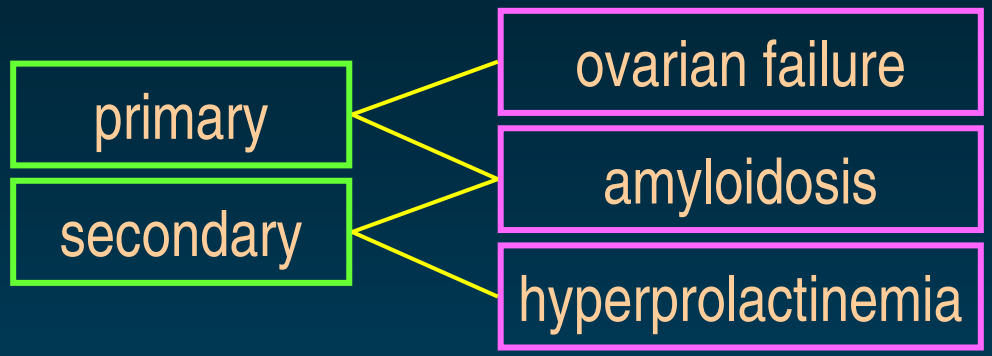
primary	lacrymal atrophy
secondary	lacrymal atrophy
primary	amyloidosis
secondary	amyloidosis

$\rightarrow$ freq(primary, secondary) = 2

(primary, secondary)

- lacrymal atrophy
- amyloidosis

Method Transforming terms



primary ovarian failure
primary amyloidosis
primary hyperprolactinemia

secondary ovarian failure
secondary amyloidosis
secondary hyperprolactinemia

Method Mapping to UMLS / SNOMED

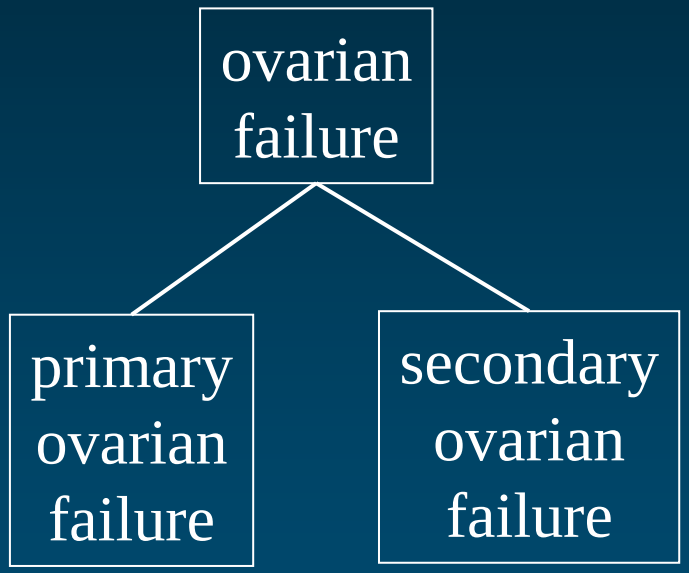
primary ovarian failure
secondary ovarian failure
primary amyloidosis
secondary amyloidosis
primary hyperprolactinemia
secondary hyperprolactinemia

UMLS
(SNOMED)

Method Analyzing the relationships

UMLS
(SNOMED)

ovarian failure

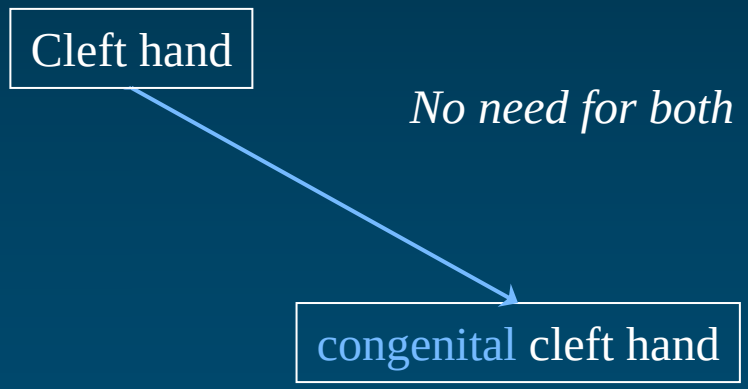


primary ovarian failure

secondary ovarian failure

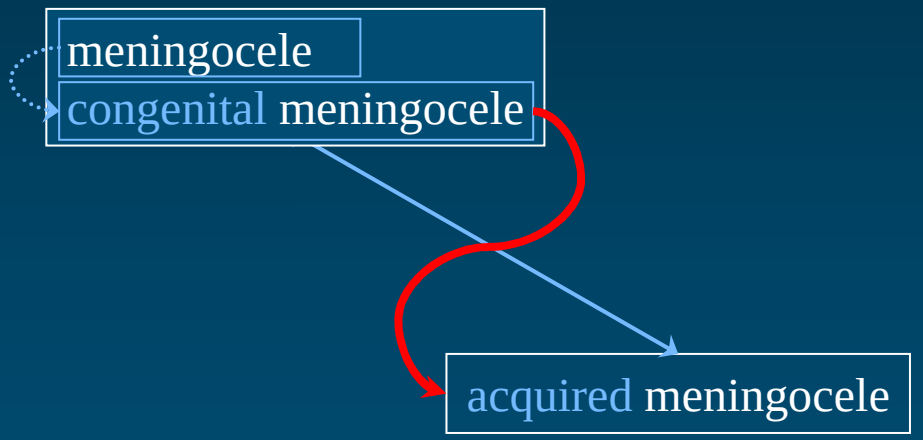
Issue 1 Missing referent

- We artificially created terms by associating modifiers with context
- Medical knowledge



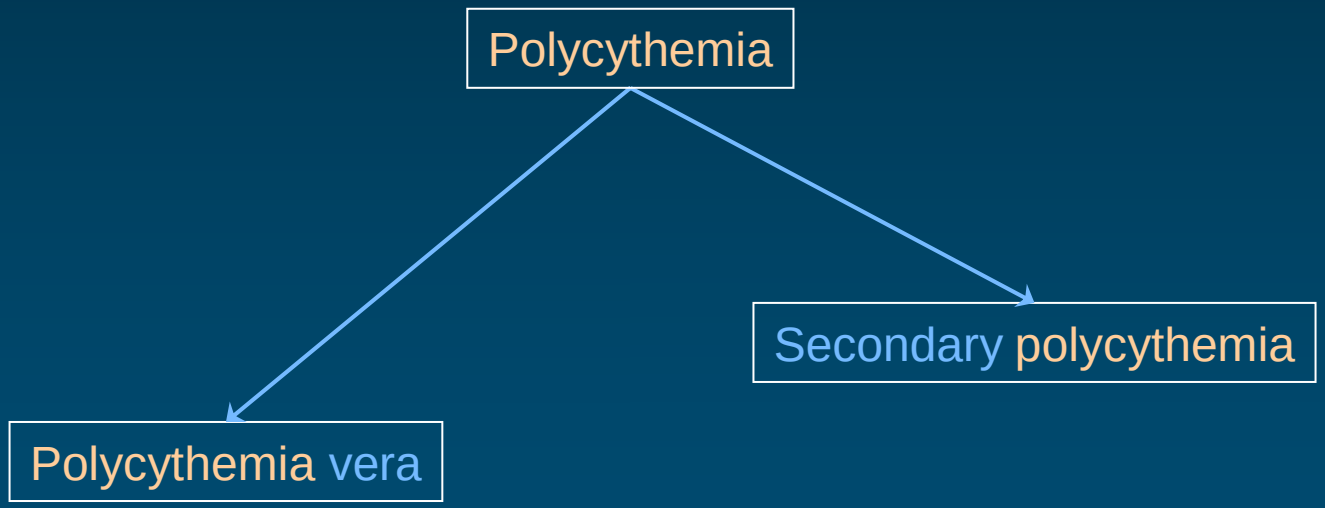
Issue 2 Missing concept

- Knowledge representation, knowledge acquisition
- Distinction among concepts
- Typical form



Issue 3 Missing symbol

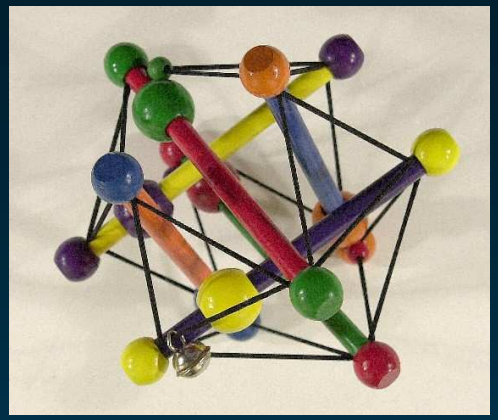
- Lexical knowledge
- Synonymy



Conclusions

Conclusion

- Use of lexical knowledge to help
  - *Build* ontologies
  - *Align* ontologies
  - *Validate* ontologies and terminologies
- These methods help automate these processes
- Domain knowledge is required



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: etbsun2.nlm.nih.gov:8000

NATIONAL LIBRARY OF MEDICINE

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA